KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

On Some New Bibliometric Applications of Statistics Related to the H-Index

Wolfgang Glänzel

DEPARTMENT OF MANAGERIAL ECONOMICS, STRATEGY AND INNOVATION (M

OR 0801

ON SOME NEW BIBLIOMETRIC APPLICATIONS OF STATISTICS RELATED TO THE H-INDEX*

Wolfgang Glänzel

Steunpunt O&O Indicatoren & K.U.Leuven, Faculty ETEW, MSI, Leuven (Belgium)

and

Hungarian Academy of Sciences, IRPS, Budapest (Hungary)

In this note some new fields of application of Hirsch-related statistics are presented. Furthermore, so far unrevealed properties of the h-index are analysed in the context of rank-frequency and extreme-value statistics.

Introduction

Since its introduction in 2005, the h-index has mainly been used as a measure to quantify the research output of individual scientists. This is in line with Jorge E. Hirsch's intensions (Hirsch, 2005). Recent attempts to fine-tune or improve the indicator (e.g., Egghe, 2006 and Jin et al., 2007) or to extend its use to higher levels of aggregation (e.g., Braun et al., 2005) follow the original design. In what follows, we will show some new application possibilities of the h-index in the context of rank statistics. In particular, the properties of the characteristic extreme values of Pareto-type distributions provide the basis of the new statistics. The first application is actually found in the form of a composite indicator strongly related to the h-index. The second application relates the h-index with a generalised version of the Zipf-Mandelbrot law. While the first indicator can only be applied to distributions with finite expectation, that is $\alpha > 1$, the second application even works if $\alpha \leq 1$. Both applications are useful supplements in evaluative studies of research performance at the micro and meso level.

Theoretical background

In recent papers (*Glänzel*, 2006, *Egghe* and *Rousseau*, 2006, *Burrell*, 2007), attempts were made to interpret theoretically some properties of the *h*-index and to connect the results with traditional indicators of publication activity and citation impact (*Braun* et al., 2005, 2006, *Schubert* and *Glänzel*, 2007). The underlying citation distribution was assumed to be Paretian and on the basis of extreme-value statistics, important properties and regularities could be derived from the distribution. Specifically, the dependence of the *h*-index on the basic parameters of the distribution and on the sample size was discussed using Gumbel's characteristic extreme values (*Gumbel*, 1958). In order to further elaborate these new approaches, we briefly summarise the mathematical rudiments.

---

* This note is based on papers presented at the 2007 Conference of the *Canadian Association for Information Science* held in Montreal (Quebec, Canada), 10-12 May 2007 and the *12th Nordic Workshop on Bibliometrics and Research Policy* (Copernhagen, Denmark), 13-14 September 2007.

Let X be a random variable. In the present case X represents the citation rate of a paper. The probability distribution of X is denoted by $p_k = P(X = k)$ for every $k \geq 0$ and the cumulative distribution function is denoted by $F_k = P(X < k)$. Put $G_k := 1 - F_k = P(X \geq k)$. Assume we have sample of size n ($\{X_i\}_{i = 1, \ldots, n}$) where all elements are independent and have the same distribution F. Gumbel's r-th characteristic extreme value (u_r) is then defined as

$$u_r := G^{-1}(r/n) = \max\ \{k: G_k \geq r/n\}\ , \tag{1}$$

where n is a given sample with distribution F (see *Gumbel*, 1958). The actual rank statistics $R(r) = X_r^*$ (where $X_1^* \geq X_2^* \geq \ldots \geq X_i^* \geq \ldots \geq X_n^*$ are the elements of the sample $\{X_i\}_{i = 1, \ldots, n}$ ranked in decreasing order) can be considered an estimator of the corresponding r-th characteristic extreme value u_r.

According to *Glänzel* (2006), the theoretical h-index (h) can be defined as

$$h := \max\ \{r: u_r \geq r\} = \max\ \{r: \max\ \{k: G_k \geq r/n\} \geq r\}\ . \tag{2}$$

If there exists such index r so that $u_r = r$ then we have obviously $h := r$ and we can write $h := u_h$.

Hirsch-related statistics for the Lomax distribution

For simplicity's sake we assume that the citation distribution under study can be approximated by a non-negative continuous distribution. In the case of continuous distributions we will write $F(x)$ and $G(x)$ instead of F_x and G_x, respectively. Furthermore, we assume that the underlying citation rates follow a Pareto distribution of the second kind. This general form of the Pareto distribution, also referred to as *Lomax* distribution, can be obtained from the infinite beta distribution if one of the parameters is chosen 1 (e.g., *Johnson, Kotz*, *Balakrishnan*, 1994). In particular, we say that the non-negative random variable X has a Pareto distribution (of the second kind) if

$$G(x) = P(X \geq x) = N^{\alpha}/(N + x)^{\alpha}\ , \text{ for all } x \geq 0 \tag{3}$$

Clearly, if x is large ($x >> N$) we can neglect the parameter in the denominator and we have

$$G(x) \sim N^{\alpha}/x^{\alpha}\ , \text{ for } x >> N. \tag{4}$$

Assuming a statistical sample with Lomax distribution and size n we obtain

$$G(u_r) \sim N^{\alpha}/u_r^{\alpha} = r/n, \text{ if } n >> r. \tag{5}$$

Consequently, we have $r \cdot u_r^{\alpha} = N^{\alpha} \cdot n$ and

$$\zeta(r) := r^{1/(\alpha+1)} \cdot u_r^{\alpha/(\alpha+1)} = N^{\alpha/(\alpha+1)} \cdot n^{1/(\alpha+1)}. \tag{6}$$

Since the right-hand side does not depend on the particular rank r, the left-hand side must be a constant. Furthermore, we have $\zeta(h) = h$ by definition (namely $\zeta(h) = h^{1/(\alpha+1)} \cdot h^{\alpha/(\alpha+1)} = h$). Consequently, $\zeta(r)$ is a constant function, moreover we have

$$\zeta(r) \equiv h \quad \text{for all } r << n. \tag{7}$$

Eq. 6 plays the central part in the following; the left-hand side will be used as the base for the analysis of the relationship with other bibliometric indicators, whereas the right-hand side provides the tools for the statistical analysis of the tails of citation distributions in the context of the h-index.

Relationship with other bibliometric indicators

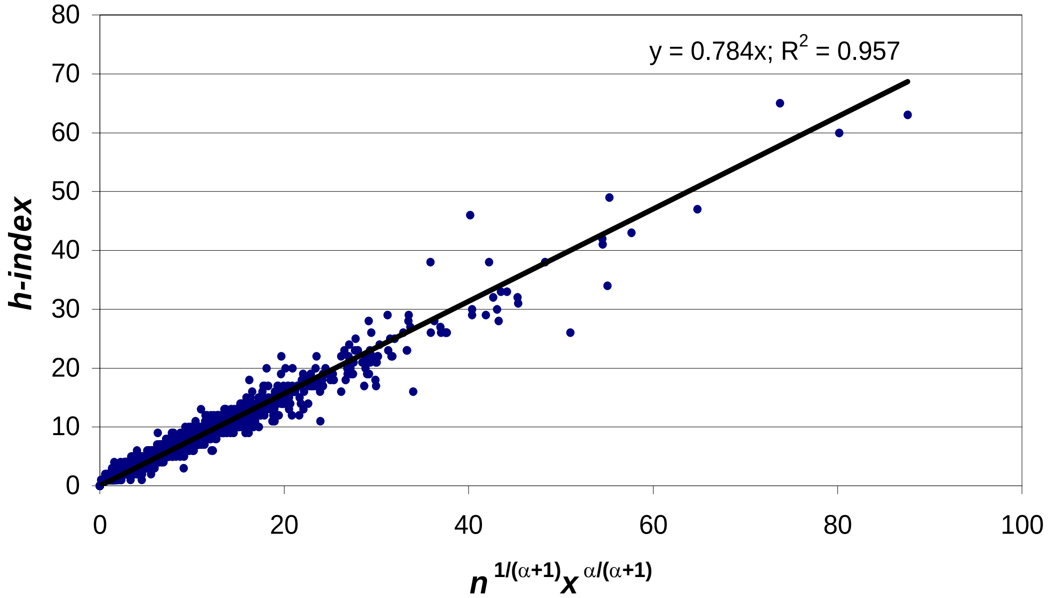
The property described in Eq. 6 and 7 yields the first important result if we take into consideration that the expected value of the Lomax distribution is $E(X) = N/(\alpha-1)$. According to *Schubert* and *Glänzel* (2007) we then have

$$h = c(\alpha)^{*} \cdot E(X)^{\alpha/(\alpha+1)} \cdot n^{1/(\alpha+1)}, \quad \text{if } \alpha > 1, \tag{8}$$

where $c(\alpha)^{*} = (\alpha-1)^{\alpha/(\alpha+1)}$ is a positive real value which only depends on the parameter α. Taking into account that the continuous *Lomax distribution* model often rather poorly fits the empirical discrete, integer-valued distributions, a perfect correlation might not be expected. Nonetheless, in their paper *Schubert* and *Glänzel* (2007) have found a strong correlation between h and $\hat{x}^{\alpha/(\alpha+1)} \cdot n^{1/(\alpha+1)}$ with $\hat{x}$ being the mean citation rate of scientific journals. Solely the empirical $c(\alpha)^{*}$ value was usually somewhat lower that the theoretical one. This result proved amazingly stable for small citation windows and it was independent of the subject field. However, the parameter α did not prove time-invariant. For largely different citation windows we have found solutions with different α values (*Glänzel*, 2007). For small windows comprising an initial period of about three years after publication, an α value around 2 has been found appropriate. For larger windows lower values yield an optimum solution. This change of exponent α with growing time intervals is in line with observations by *Vlachý* (1976) and *Pao* (1986). This effect is shown on an example for journals. Figure 1 shows the dependence of h on n and journal impact measures $\hat{x}$ for papers published in 1980 and indexed in the *Science Citation Index* of Thomson Scientific (Philadelphia, PA, USA). The impact measures have been calculated for a 3-year (top) and 21-year (bottom) citation window, respectively. In the first case $\alpha = 2$, for the longer citation period $\alpha = 1.5$ has been chosen.

Both theoretical considerations and empirical analysis therefore lead to the conclusion that the h-index strongly correlates with $\hat{x}^{\alpha/(\alpha+1)} \cdot n^{1/(\alpha+1)}$ which can be considered a composite indicator combining publication output and mean citation rate. Although such composite indicator has interesting properties, it is not intended to substitute the h-index. In this context we mention that a similar composite indicator for the journal impact was already suggested by *Lindsey* (1978) independently from the Hirsch-index theory. His idea was introducing a measure that takes into account the overall quality of production and which can be used in cross-disciplinary studies, too. In particular he used the geometric mean of the number of publications P and the received citations C as balancing correction factor for the mean citation rate. Thus his *Corrected Quality*

Ratio (CQ) is defined as CQ = $(C/P)\cdot(C\cdot P)^{½} = (C^3/P)^{½}$. Applying our notation we actually have CQ = $n\cdot\hat{x}^{3/2}$. By applying the transformation CQ $\longmapsto$ $\mathrm{CQ}^{0.4}$ we obtain $\mathrm{CQ}^{0.4} = n^{0.4}\cdot\hat{x}^{0.6}$, which coincides with the above h-based composite indicator for the choice $\alpha = 1.5$. Since the power function is a strictly monotonous function, the applied transformation has no effect on the ranking according to Lindsey's Corrected Quality Ratio.



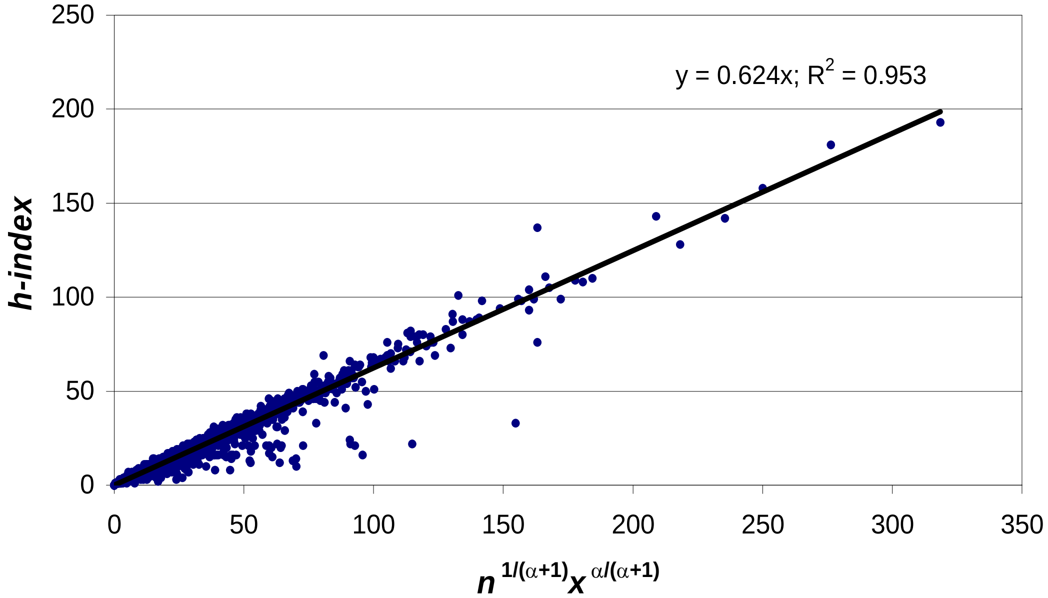


Figure 1 Correlation of the journal h-index with $n^{1/(\alpha+1)}\cdot\hat{x}^{\alpha/(\alpha+1)}$ in all science fields combined top: citation window: 1980-1982 (α=2); bottom: citation window: 1980–2000 (α=1.5)

The z(r) statistics and their properties

A second important regularity can be obtained when replacing the theoretical values in the left-hand side of Eq. 6 by the corresponding statistics, that is, the Gumbel extreme values by the corresponding ranked sample elements. In particular, we define

$$z(r) := r^A \cdot R(r)^{(1-A)} \text{ with } A = 1/(\alpha+1), \quad (9)$$

where $z(r)$ is expected to be an estimator of the expression in right-hand side of Eq. 6 and thus of h for each $r << n$.

Unfortunately, z is not an unbiased estimator of h but the following properties hold for certain transformations of $z(r)$. In particular, we analyse the statistics $k \cdot \ln(z(r)/z(r+1))$ for $r < n$.

Proposition 1: The $k \cdot \ln(z(r)/z(r+1))$ statistics are independent (non identically) exponentially distributed random variables with expectation

$$\mathrm{E}\{r \cdot \ln[z(r)/z(r+1)]\} = \{r \cdot \ln[r/(r+1)] + 1\}/(\alpha+1) \qquad (10)$$

which tends to 0 if $r \to \infty$.

Proof: The proof is a consequence of a theorem by *Glänzel* et al. (1984) taking into account that

$$r \cdot \ln[z(r)/z(r+1)] = \{r \cdot \ln[r/(r+1)] + \alpha \cdot r \cdot \ln(R(r)/R(r+1))\}/(\alpha+1)$$

and $\alpha \cdot r \cdot \ln(R(r)/R(r+1))$ are independent identically distributed random variables, where the common distribution is exponential with parameter 1. + + +

Proposition 2: The expected value of $\mathrm{mean}_m(r \cdot \ln(z(r)/z(r+1)))$ tends to 0 as $m,n \to \infty$ and for its standard deviation we have

$$\mathrm{D}\left[\mathrm{mean}_m(r \cdot \ln[z(r)/z(r+1)])\right] = \mathrm{D}\left[\frac{1}{m} \cdot \sum_{r=1}^{m} r \cdot \ln \frac{z(r)}{z(r+1)}\right] = m^{-½}/(\alpha+1)$$

for all $m < n$.

Proof: First we define the following expression for any $m < n$:

$$Z(m) := (\alpha+1) \cdot \mathrm{E}\left[\frac{1}{m} \cdot \sum_{r=1}^{m} r \cdot \ln \frac{z(r)}{z(r+1)}\right]$$

Elementary manipulations result in the following equation.

$$Z(m) := \frac{1}{m} \cdot \mathrm{E}\left[\sum_{r=1}^{m} r \cdot \ln \frac{r}{r+1} + \alpha \sum_{r=1}^{m} r \cdot \ln \frac{R(r)}{R(r)+1}\right] =$$

$$= \frac{1}{m} \cdot \sum_{r=1}^{m} \ln \frac{r^r}{(r+1)^r} + \frac{1}{m} \cdot \mathrm{E}\left[\alpha \sum_{r=1}^{m} r \cdot \ln \frac{R(r)}{R(r)+1} \right]$$

Note that E[$\alpha \cdot r \cdot \ln(R(r)/R(r+1))$] = 1 independently of r because of the underlying Pareto distribution (cf. proof of Proposition 1). Therefore the second expression of the right-hand side takes the value 1 for any m. Applying Stirling's approximation for the factorial function to the first expression we obtain

$$Z(m) := \ln \frac{(m+1)!}{(m+1)^{m+1}} + 1 \sim \frac{0.5 \ln(m+1) - 0.081}{m} \to 0 \text{ as } m \to \infty.$$

The same applies to $\mathrm{E}[\mathrm{mean}_m(r \cdot \ln(z(r)/z(r+1)))] = Z(m)/(\alpha+1)$ with $\alpha > 0$. This completes the first part of the proof.

The proof of the second part is in principle straightforward since $\mathrm{mean}_m(r \cdot \ln[z(r)/z(r+1))$ is the sum of different linear combinations of the independent identically distributed random variables $r \cdot \ln(R(r)/R(r+1))$. In the following we give, however, a more detailed proof. We use the following notations $c_r := r \cdot \ln(r/(r+1))$ and $x_r := \alpha \cdot r \cdot \ln(R(r)/R(r+1))$. Thus we have

$$(\alpha+1)^{-2}\mathrm{D}^2[\mathrm{mean}_m(r \cdot \ln[z(r)/z(r+1))] =$$

$$= \mathrm{E}[\{\Sigma(c_r+x_r)\}^2]/m^2 - \{\mathrm{E}[\Sigma(c_r+x_r)/m]\}^2 =$$

$$= \mathrm{E}[(\Sigma c_r)^2 + 2\Sigma c_r \Sigma x_r + (\Sigma x_r)^2]/m^2 - (\Sigma c_r + m)^2/m^2 =$$

$$= \{(\Sigma c_r)^2 + 2m\Sigma c_r + \mathrm{E}(\Sigma x_r)^2 - (\Sigma c_r + m)^2\}/m^2 =$$

$$= \{(\Sigma c_r)^2 + 2m\Sigma c_r + \Sigma \mathrm{E}[x_r^2] + 2\Sigma_{r \neq s} \mathrm{E}[x_r \cdot x_s] - (\Sigma c_r + m)^2\}/m^2 =$$

$$= \{(\Sigma c_r)^2 + 2m\Sigma c_r + \Sigma \mathrm{E}[x_r^2] + 2\Sigma_{r \neq s} \mathrm{E}[x_r \cdot x_s] - (\Sigma c_r + m)^2\}/m^2 =$$

$$= \{\Sigma \mathrm{E}[x_r^2] + 2\Sigma_{r \neq s} \mathrm{E}x_r \cdot \mathrm{E}x_s - m^2\}/m^2$$

Since x_r has an exponential distribution with parameter 1, we have $\mathrm{E}x_r = 1$ and $\mathrm{E}[x_r^2] = 2$. Finally we have

$$(\alpha+1)^{-2}\mathrm{D}^2[\mathrm{mean}_m(r \cdot \ln[z(r)/z(r+1))] = \{2m + m(m-1) - m^2\}/m^2 = 1/m.$$

This completes the proof. + + +

Corollary: For the mean of the *Hirsch core* we have $m = h$ and consequently

$$Z(h) := \ln \frac{(h+1)!}{(h+1)^{h+1}} + 1 \sim \frac{0.5 \ln(h+1) - 0.081}{h} \sim 0, \text{ provided } h \text{ is large enough.}$$

The corresponding Z values for some h values are presented in Table 1. Note that $\mathrm{mean}_h(r \cdot \ln[z(r)/z(r+1))$ is an unbiased estimator of $Z(h)/(\alpha+1)$. For $\alpha = 2$, for instance, the Z values must be divided by 3, and the corresponding expectations therefore range between 0.038 for h = 10 and 0.001 for h = 1000.

Table 1 h and $Z(h)$ values for different orders of magnitude for h

h	10	25	50	100	500	1000
Z(h)	0.113	0.062	0.037	0.022	0.006	0.003

Since the exponential distribution belongs to the domain of attraction of the normal distribution, we can apply a Welsch-test to the mean values provided that the size of the underlying paper set amounts to about 25 or more (see *Schubert* and *Glänzel*, 1983). The statistic $w = (\alpha+1)\cdot h^{1/2}\{\text{mean}_h(r\cdot\ln[z(r)/z(r+1)) - Z(h)/(\alpha+1)\}$ has approximately a standard normal distribution. Thus, if $|w| < w^*$, the null-hypothesis H_0: $\text{mean}_h(r\cdot\ln[z(r)/z(r+1)) = Z(h)/(\alpha+1)$ can be accepted at a significance level of $2\Phi(w^*)-1$, where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. This test can be used an *indirect* goodness-of-fit test for the h-property of the tail of the empirical distribution. For the significance level of 0.95 we have w^* = 1.96. Assuming an h-index of 25 and a small citation window with α = 2, we obtain the following confidence interval for $\text{mean}_h(r\cdot\ln[z(r)/z(r+1))$. We can accept the h-property for the z statistics if $\text{mean}_h(r\cdot\ln[z(r)/z(r+1))\in(-0.100, 0.162)$. For h=100, for instance, we have $\text{mean}_h(r\cdot\ln[z(r)/z(r+1))\in(-0.058, 0.072)$ according to the values given in Table 1.

This new method for the analysis of the tail properties of Pareto-type distributions based on the z statistics works much better than an earlier model described by *Glänzel* and *Schubert* (1988). The latter one was based on transformations of ordered statistics, namely on individual $r\cdot\ln(X_r^*/X_{r+1}^*) = r\cdot\ln[R(r)/R(r+1)]$ statistics with $r \ll n$, which were extremely sensitive to ties. In practice, rank statistics of integer-valued discrete distributions often include ties (i.e. $R(r) = R(r+1)$ for some r = 1, 2, …) resulting in $r\cdot\ln[R(r)/R(r+1)] = 0$. These ties can heavily distort the fit of the exponential distribution and the applied goodness-of-fit tests. By contrast, the new z statistics are more robust and much less sensitive to ties. In order to give an example, we have calculated the h-index, the R and z statistics as well as the above means for eight high impact journals from different fields. The publication year was 1980 and the citation window embraces three years beginning with the year of publication. Table 2 presents the corresponding statistics around the h value. Although the deviation from $\zeta(r) \equiv h$ is quite large for the individual r values but the median M of the empirical $z(r) = r^A\cdot R(r)^{(1-A)}$ values and the means introduced above provide strikingly robust estimators of h.

Table 2 h-related statistics of eight selected journals (publication year: 1980, citation window: 1980-1982)

Science			Nature			Cell			Blood		
r	*R*(r)	*z*(*r*)	*r*	*R*(r)	*z*(*r*)	*r*	*R*(r)	*z*(*r*)	*r*	*R*(r)	*z*(*r*)
97	112	106.8	99	113	108.1	69	81	76.8	41	54	49.3
98	111	106.5	100	112	107.8	70	81	77.2	42	53	49.0
99	111	106.8	101	111	107.6	71	81	77.5	43	53	49.4
100	111	107.2	102	110	107.3	72	81	77.9	44	53	49.8
101	110	106.9	103	110	107.6	73	80	77.6	45	53	50.2
102	110	107.3	104	109	107.3	74	79	77.3	46	52	49.9
103	110	107.6	105	109	107.7	75	79	77.6	47	52	50.3
104	109	107.3	106	109	108.0	76	79	78.0	48	52	50.6
105	106	105.7	107	108	107.7	77	79	78.3	49	51	50.3
106	**106**	**106.0**	**108**	**108**	**108.0**	**78**	**79**	**78.7**	**50**	**50**	**50.0**
107	106	106.3	109	108	108.3	79	78	78.3	51	49	49.7

Angewandte Chemie			Astrophysical Journal			Analytical Chemistry			Trends in Neurosciences		
r	*R*(r)	*z*(*r*)	*r*	*R*(r)	*z*(*r*)	*r*	*R*(r)	*z*(*r*)	*r*	*R*(r)	*z*(*r*)
34	49	43.4	40	52	47.6	23	40	33.3	15	35	26.4
35	48	43.2	41	52	48.0	24	38	32.6	16	34	26.4
36	48	43.6	42	52	48.4	25	38	33.0	17	32	25.9
37	47	43.4	43	52	48.8	26	36	32.3	18	32	26.4
38	46	43.2	44	51	48.6	27	36	32.7	19	29	25.2
39	46	43.5	45	50	48.3	28	36	33.1	20	28	25.0
40	44	42.6	46	50	48.6	29	34	32.2	21	25	23.6
41	43	42.3	47	50	49.0	30	34	32.6	22	25	24.0
42	43	42.7	48	49	48.7	31	34	33.0	23	25	24.3
43	**43**	**43.0**	**49**	**49**	**49.0**	**32**	**33**	**32.7**	**24**	**25**	**24.7**
44	42	42.7	50	49	49.3	33	32	32.3	25	23	23.6

The medians y of the estimates of h and the means of the $r \cdot \ln[z(r)/z(r+1)]$ statistics for the eight journals are given in Table 3. None of the means exceeds their critical values belonging to the significance level of 0.95. Thus the statistics based on the empirical journal samples reflect sufficiently well the h-property of the tail of citation distributions described by the right-hand side of Eq. 6. At the same time we can conclude that h can be used as an appropriate truncation point for the tail of a distribution.

Table 3 Basic statistics for the analysis of the *Hirsch core*

Journal	h-index	$y = median(z(r))$	$mean_h\{r \cdot \ln[z(r)/z(r+1)]\}$
Science	106	104.5	-0.036
Nature	108	110.0	+0.021
Cell	78	78.0	-0.038
Blood	50	49.5	-0.053
Angewandte Chemie	43	43.6	+0.017
Astrophysical Journal	49	47.1	-0.056
Analytical Chemistry	32	32.0	-0.021
Trends in Neurosciences	24	24.5	-0.026

Conclusions

In this paper we have described two new applications of Hirsch-related indexes. The composite indicator, which expresses a multiplicative connection between derivatives of publication output and citation impact, proved surprisingly robust and works at both the meso and the micro level. Its strong correlation with the h-index is independent of the subject area (cf. *Schubert* and *Glänzel*, 2007). The *z* statistics, representing the second application, can be used to analyse the tail of citation distributions in the light of the h-index. At the same time, the h-index proved useful as truncation point for rank frequency analysis, for instance, by applying *z* and related statistics to the *Hirsch core* (e.g. *Burrell*, 2007) publications.
Finally, a further important property is worth mentioning in this context, namely that the *z* statistics can be considered a version of the Zipf-Mandelbrot law (cf. *Yablonski*, 1980, *Egghe* and *Rousseau*, 1990), where the constant value equals the h-index to the power $(\alpha+1)$, that is, $r\cdot R(r)^{\alpha} = \{z(r)\}^{\alpha+1} = h^{\alpha+1}$. In particular, the special case $\alpha = 1$, which corresponds to the Lotka distribution, results in the following version of the classical Zipf's Law: $z(r) = \{r\cdot R(r)\}^{½} = h$ = constant, or equivalently, $r\cdot R(r) = h^2$.

References

Braun, T., Glänzel, W., Schubert, A. (2005), A Hirsch-type index for journals. *The Scientist*, 19 (22) 8.

Braun, T., Glänzel, W., Schubert, A. (2006), A Hirsch-type index for journals, *Scientometrics*, 69 (1), 169-173

Burrell, Q. L. (2006), Hirsch's h-index: a stochastic model. *Journal of Informetrics*, 1 (1) 16-25

Burrell, Q. L. (2007), On the h-index, the size of the Hirsch core and Jin's A-index, *Journal of Informetrics*, 1 (2) 170-177.

Egghe, L., Rousseau, R. (1990). *Introduction to informetrics. Quantitative methods in library, documentation and information science*. Elsevier Science Publisher. Amsterdam.

Egghe, L., Rousseau, R. (2006). An informetric model for the h-index. Scientometrics, 69(1), 121-129.

Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics*, 69(1), 131-152.

Glänzel, W., Schubert, A., Telcs, A. (1984), *Goodness of Fit Test for the Tail of Distributions*. Bolyai Colloquium on Goodness of fit (Debrecen, Hungary, June 25-28, 1984).

Glänzel, W., Schubert, A. (1988), *Theoretical and Empirical Studies of the Tail of Scientometric Distributions*. In: L. Egghe, R. Rousseau (Eds.), Informetrics 87/88, Elsevier Science Publisher, 75-83.

Glänzel, W. (2006), On the h-index – A mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, 67 (2) 315–321.

Glänzel, W. (2007), Some new applications of the h-index, *ISSI Newsletter*, 3 (2), 28-31.

Gumbel, E. J. (1958). *Statistics of extremes*. New York: Columbia University Press.

Hirsch, J. E. (2005), An index to quantify an individual's scientific research output, Proceedings of the National Academy of Sciences of the United States of America, 102 (46) 16569–16572. (also available at: arXiv:physics/0508025, accessible via http://arxiv.org/abs/physics/0508025).

Jin, B.H., Liang, L.M., Rousseau., R., Egghe, L. (2007), The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, 52 (6) 855-863.

Johnson, N. L., Kotz, S., Balakrishnan, N. (1994), *Continuous univariate distributions*. Volume 1, 2nd Edition, John Wiley & Sons, Ney York.

Schubert, A. Glänzel, W. (1983), Statistical reliability of comparisons based on the citation impact of scientific publications. *Scientometrics*, 5 (1) 59-74.

Schubert, A. Glänzel, W. (2007), A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics*, 1 (3), in press. (doi:10.1016/j.joi.2006.12.002)

Pao, M. L. (1986), An empirical examination of Lotka's law. *Journal of the American Society for Information Science*, 37 (1) 26–33.

Vlachý, J. (1976), Time factor in Lotka's law. *Probleme de Informare si Documentare*, 10 (2) 44–87.

Yablonski, A. I. (1980), On fundamental regularities of the distribution of scientific productivity. *Scientometrics*, 2 (1) 3-34.